12-2023

A Cancer-Specific Study on the Differentially Expressed Protein-Protein Interactions of Fumarate Hydratase

Sydney Lac

A Cancer-Specific Study on the Differentially Expressed Protein-Protein Interactions of Fumarate Hydratase

by

Sydney Lac, BSA

APPROVED:

DocuSigned by:
Xiaobo Zhou
185774089A554BC...
Xiaobo Zhou, Ph.D.
Advisory Professor

DocuSigned by:
pora kim
4A5E60F1D48A44C...
Pora Kim, Ph.D.

DocuSigned by:
Nami McCarty
A31617AA0FC946D...
Nami McCarty, Ph.D.

DocuSigned by:
Wenjin Zheng
6B723716D187436...
Wenjin Zheng, Ph.D.

DocuSigned by:
Nidhi Sahni
79674F8A2E8B460...
Nidhi Sahni, Ph.D.

APPROVED:

____
Dean, The University of Texas
MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Sciences

A Cancer-Specific Study on the Differentially Expressed Protein-Protein Interactions of Fumarate Hydratase

A

Thesis

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth Houston

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

by

Sydney Lac, BSA
Houston, Texas

December, 2023

A Cancer-Specific Study on the Differentially Expressed Protein-Protein Interactions of Fumarate Hydratase

Sydney Lac, BSA

Advisory Professor: Xiaobo Zhou, Ph. D.

Fumarate hydratase (FH) is an enzyme used in the Krebs Cycle to convert fumarate to malate, and it is controlled by the FH gene. In this paper, we will investigate its role in Uterine Corpus Endometrial Carcinoma (UCEC) and how FH-deficient cells affect tumorigenesis. It is well-established that FH has been extensively studied in connection with renal cell carcinoma, skin and uterine leiomyomas, pheochromocytoma, and paraganglioma. However, we aim to construct an interaction network of significant genes related to the FH gene under conditions of FH deficiency in the Kreb Cycle. Creating an interactive network that illustrates the interconnectedness of FH's role is crucial for comprehending cellular adaptations when FH is deficient. Unfortunately, we have not yet found a reliable and accurate representation of this complex network, which has prompted us to create our own. For our dataset, we utilized RNAseq count data from the UCSC Xena database. We followed this with a differential expression gene (DEG) analysis workflow involving Limma and EdgeR. The significantly expressed genes were contextualized through an enrichment analysis called EnrichGO. Finally, we associated the significantly expressed genes with a transcription factor (TF). Our results have allowed us to construct a network that presents our findings. Most importantly, it has revealed the significant role played by the HIF3a TF in FH-deficient cells. While HIF3a is less understood compared to its other isoforms (HIF1a and HIF2a), this research contributes to bridging that knowledge gap. Our findings suggest that the HIF3a gene is a significantly differentially expressed gene in FH deficient patients.

Table of Contents

List of Figures

List of Tables

Introduction

Fumarate hydratase is an enzyme used in the Krebs Cycle to convert fumarate to malate. Uterine Corpus Endometrial Carcinoma (UCEC) is a cancer type in which FH mutations are notably prevalent, as reported by Goldman et al. in 2020. Notably, FH has been the subject of numerous studies linking it to various health conditions, including renal cell carcinoma, skin and uterine leiomyomas, pheochromocytoma, and paraganglioma, as documented by Fuchs et al. in 2023 and Toro JR et al. in 2003. These studies have explored the role of FH in causing or being associated with these specific diseases. Our objective is to construct an interaction network that highlights significant genes related to the FH gene under conditions of FH deficiency in the Kreb Cycle. The creation of an interactive network that portrays the interconnectedness of FH's role is vital for comprehending how our cells adapt when FH is deficient.

Proteins are the fundamental building blocks responsible for the body's regeneration, communication, and the maintenance of homeostasis. To gain a deeper understanding of how a gene can be linked to cancer, it is essential to establish a solid foundation of basic interaction information. This foundation includes understanding how a gene interacts with and influences multiple pathways within the body. For instance, FH is known to have effects on the Urea cycle, DNA translation, the glucose cycle, and more, as detailed by Schmidt et al. in 2020. However, we have not yet found a dependable and accurate representation of this complex network. As a result, we are creating our own representation.

In addition to the gene network, we will incorporate the transcription factors associated with those genes in the network. Transcription factors are crucial proteins involved in the regulation of gene expression within cells, as discussed by Wang et al. in 2015. Their functions include modulating gene expression, both activation and repression, and aiding in cell type specificity.

In this study, we will examine significant differentially expressed genes in UCEC patients. To ensure comprehensive coverage of FH, we will explore various databases available online to further develop the data we have acquired. Additionally, we aim to determine the most effective way to present our findings for user-friendliness while encompassing all relevant information. Upon the project's completion, we anticipate a deeper understanding of FH and its functions throughout the cell. The primary objective of this study is to construct and project a gene interaction map for fumarate hydratase, enhancing our comprehension of its role in the Kreb Cycle. Our hypothesis suggests that the creation of this gene interaction map will facilitate a clearer understanding of FH's cellular processes and how a disease state affects the cell.

Methods

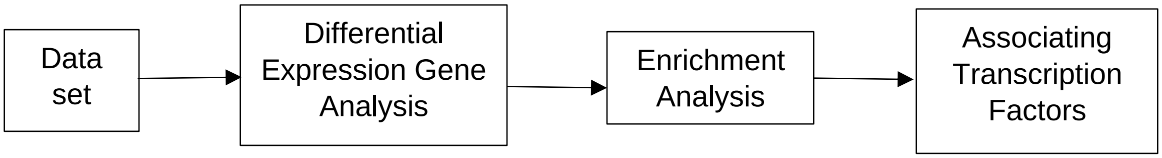

Data set
Differential Expression Gene Analysis
Enrichment Analysis
Associating Transcription Factors

Figure 1: Overall Workflow

Overall Workflow

We initiated our study by exploring the cBioPortal database. Our primary focus was the gene of interest, FH, and we observed that UCEC had the highest number of mutated patients among cancer types. Subsequently, we retrieved the count data from the University of

California Santa Cruz (UCSC) Xena database. Initially, this dataset encompassed 583 patients, but 35 were excluded due to the absence of mutations in either FH or the UCEC cancer type. Using the adjusted dataset, we performed a Differential Expression Gene Analysis (DEG) using two methods in R Studio: Limma and EdgeR. The result files obtained from Limma and EdgeR were filtered based on a minimum threshold for p-adjusted values and log2foldchange values. The differentially expressed genes (DEGenes) that passed the threshold were ‘significant’. Subsequently, we compared and overlapped the results from both methods, designating those DEGenes also as 'significant.' These significant DEGenes progressed to the next stage of our workflow, which involved Enrichment Analysis. We employed the EnrichGO method in R Studio for this analysis. EnrichGO is a Gene Ontology (GO) term enrichment platform that categorizes genes based on their functional attributes. The final step in our workflow entailed associating transcription factors (TFs) with the DEGenes. To achieve this, we utilized a web-based tool called ChEA3, which predicts the TFs associated with the inputted list of DEGenes. It employs Fisher's Exact Test to compare the input list of genes with its TF database to identify the closest associations. Additionally, we created an illustrative graphic to provide an overview of the relationship between FH, DEGenes, and transcription factors.

Dataset

This research paper was conducted computationally, and our initial step involved selecting the appropriate study population. We utilized the cBioPortal for Cancer Genomics tool (Cerami et al., 2012) to facilitate our research. In the 'Quick Search' bar, we queried for the FH gene, drawing data from the TCGA Pan-Cancer studies. From the bar graphic that cBioPortal generates, we chose the cancer type with the most amount of ‘mutated’ FH.

Among the informative graphics provided by cBioPortal, one particularly valuable visual aided in identifying the specific locations within the FH gene where mutations occurred was called a lollipop graph. Our criterion for the minimum number of mutated FH patients was set at a minimum of 10 patients. Within the UCEC cohort, there were a total of 583 patients, of which 21 had mutations in FH. With the appropriate cancer type selected, the next step was to obtain the actual gene expression RNASeq - HTSeq counts data. We sourced this data from the UCSC Xena database (Goldman et al., 2020). The dataset included RNAseq gene expression counts presented in log2(count + 1) units and was sequenced using the Illumina platform. The cohort under examination was the GDC (Genomic Data Commons) TCGA Endometrioid Cancer (UCEC) (Goldman et al., 2020), with the dataset ID designated as TCGA-UCEC.htseq_counts.tsv (Goldman et al., 2020). The downloaded file contained records for 583 female patients and a total of 60,489 identifiers (genes), each identified via Ensembl Gene IDs. We excluded 35 patients from the dataset due to the absence of FH mutations and their non-affiliation with the UCEC cancer type. Consequently, the final dataset comprised 548 patients in total, with 21 patients exhibiting FH mutations. The comparison group for this dataset consisted of patients with mutated FH versus those without FH mutations.

Differential Expression Gene Analysis

Utilizing the dataset described in the previous paragraph, we employed two R packages in R Studio: Limma and EdgeR. Limma, short for 'linear models for microarray data,' is an R/Bioconductor software package renowned for its utility (Ritchie et al., 2015). We opted for Limma due to its popularity in differential gene expression workflows and its frequent application within our research laboratory. Limma operates by interpreting experimental values collectively, as opposed to isolating individual comparisons, such as between mutated and normal patients (Phipson et al., 2016). This method employs a statistical model known as parametric empirical Bayes, which effectively leverages the relationships between genes to control for residual variances (Law et al., 2014). Limma requires two input files: CountData, which includes gene names and sample IDs, and ColumnData, which consists of sample IDs and patient status (mutated or non-mutated). It's essential for the order of patients to be consistent in both files. Limma features its own normalization function called 'voom,' which is seamlessly integrated into the workflow (Law et al., 2014). The results file generated by Limma was subdivided into up and down-regulated genes. Up-regulated genes were defined by a Log2FoldChange greater than 1.58 and a p-value less than 0.05, while down-regulated genes were characterized by a Log2FoldChange less than -1.58 and a p-value less than 0.05. The exported file from Limma encompasses specific column headers, including logFC, AveExpr, t, P.Value, adj.P.Val, and B. Table 1 offers a concise summary of the significance of each of these headers.

Table 1: Limma Export File Headers (Ritchie et al., 2015)

Header	**Meaning**
logFC	Log2fold; gives the value of the contrast between 2 or more experimental conditions
AveExpr	Gives the average log2FC expression level for that gene across all the arrays and channels in the experiment
t	Is the moderated t-statistic
P.Value	Is the associated p-value
Adj.P.Val	Is the p-value adjusted for multiple testing
B	Is the log-odds that the gene is differentially expressed.

The subsequent tool in our analysis was EdgeR (Empirical Analysis of Digital Gene Expression Data), another Bioconductor software package that we applied within R Studio (Robinson et al., 2010). EdgeR employs the Exact Test as its statistical framework to aid in identifying differentially expressed genes. While Limma utilizes the 'voom' function for normalization, EdgeR relies on the Trimmed Mean of M values (TMM) for this purpose (Robinson & Oshlack, 2010). EdgeR utilizes the same input files as Limma, namely CountData and ColData. Additionally, the cutoff criteria for identifying up-regulated and down-regulated genes remain consistent (Log2FoldChange > 1.58 and p-value < 0.05). The exported file from EdgeR comprises column headers such as logFC, logCPM, PValue, and FDR. A concise summary of the significance of each of these column headers is provided in Table 2.

Table 2: EdgeR Export File Headers (Chen et al., 2016)

Header	**Meaning**
logFC	Log2FC between the groups
logCPM	The average log2 counts per million
PValue	The listed p-value
FDR	Adjusted p-value

The subsequent step involved the intersection of differentially expressed genes derived from Limma and EdgeR in R Studio. This intersection was performed to demonstrate the consistency of results between both tools and to ensure the inclusion of the most significantly differentially expressed genes from each. The process began by segregating up-regulated and down-regulated genes from the result files of both Limma and EdgeR. The 'intersect' command in R Studio was employed to compare up-regulated genes from Limma and EdgeR, followed by a similar comparison for down-regulated genes. Once these comparisons were completed, all intersected genes were consolidated into a single file. With this final file containing commonly identified DEGenes from both methods, we generated a heatmap to visually represent the data and observe any evident clustering. Additionally, this intersected genes file served as the foundation for the subsequent step: enrichment analysis.

Enrichment Analysis

For the enrichment analysis, we utilized the 'clusterProfiler' package in R Studio. Specifically, within the clusterProfiler package, we employed the enrichGO library to conduct the enrichment analysis. This powerful tool draws upon data from the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases (Yu et al., 2012). The Gene Ontology categorizes the input gene list into three distinct annotations: molecular functions (MF), biological processes (BP), and cellular components (CC) (Ashburner et al., 2000). As elucidated by Balakrishnan et al. (2013), MF encompasses elemental activities such as catalysis or binding, BP refers to processes specific to the functioning of living units like cells or tissues, and CC relates to the cellular location where a gene product is situated. The 'EnrichGO' method effectively controls for a high false discovery rate by estimating q-values as well(Yu et al., 2012). These q-values are included in the results file exported from R Studio.

The input for enrichGO consisted of the DEGenes derived from both the intersected Limma/EdgeR results and the EdgeR only workflow, using the gene symbols instead of Ensembl IDs. The annotation package that enrichGO accessed is known as 'org.Hs.eg.db' (Yu et al., 2012). After running enrichGO, the output file contains information about the category to which the gene(s) were assigned and the corresponding descriptive term for the GO term. In Table 3, you can find a summary of the headers for the exported enrichGO file.

Table 3: EnrichGO Export File Headers (T Wu et al., 2021)

Header	**Meaning**
Ontology	The enrichment term that helps to categorize the ID into functional characteristics (BP, CC, MF)
ID	Unique identifier for Gene Ontology

Description	Describes the ID
GeneRatio	Ratio of input genes that are annotated in a term
BgRatio	Ratio of all genes that are annotated in this term
pvalue	The stated p-value
p.adjust	Adjusted p-value
qvalue	q-value cutoff on enrichment tests to be significant
geneID	These are names of the genes part of the ID from the input list of DEGenes
Count	The number of genes in that geneID; Count = geneID

Associating Transcription Factors (TFs)

Associating transcription factors is a crucial step as these proteins are responsible for regulating the differentially expressed genes. We employed ChEA3 to carry out this step in the workflow. ChEA3, short for ChIP-X Enrichment Analysis 3, ranks TFs based on the input list we submitted (Keenan et al., 2019). The tool draws information from a database that includes TF-gene co-expression from RNAseq studies, TF-target associations from ChIPseq experiments, and TF-gene co-occurrences from crowd-sourced gene lists (Keenan et al., 2019). In this phase of the workflow, we utilized the DEGenes obtained from edgeR and input them into the ChEA3 web tool. The website populated and ranked the TF it believes is most associated with the input DEGenes. Moreover, it provides links to relevant websites and information about that TF for user convenience. We selected the top-ranked TF and exported the overlapping genes. Using this list, we compared it with the genes in the GO terms file to determine the pathway in which that TF is involved. With the DEGene, TF, and GO term/pathway, we established network connections between FH and the DEGene. A concise summary of the significance of each of these column headers is provided in Table 4.

Table 4: ChEA3 Export File Headers (Keenan et al., 2019)	
Header	**Meaning**
Query Name	The name of the query that the user inputs
Rank	Ranks the top TF based on Score

TF	Transcription Factor
Score	Indicates relevancy to the TF
Library	Pre-generated TF-target gene set libraries
Overlapping Genes	The genes that appear linked to the TF based on the input list

Results

Differential Expression Gene Analysis

After running Limma, we identified 151 up-regulated genes and 99 down-regulated genes. Subsequently, running EdgeR revealed 116 up-regulated genes and 836 down-regulated genes. Upon overlapping the results from Limma and EdgeR, we found 29 overlapping up-regulated genes and 68 overlapping down-regulated genes. A summary of these results is presented in Table 5. It's worth noting that the cut-off for all DEGenes across all methods was a log2FoldChange greater than or less than 1.58 and an adjusted p-value less than 0.05.

Table 5: Summarizing DEGene Analysis

Method	Up-regulated Genes	Down-regulated Genes	Total DEGenes
Limma only	151	99	250
EdgeR only	116	836	952
Intersected method	29	68	97

We also created a heatmap from the intersected genes file (Figure 2). From the intersected genes file, we were able to move on to the next step which was enrichment analysis.

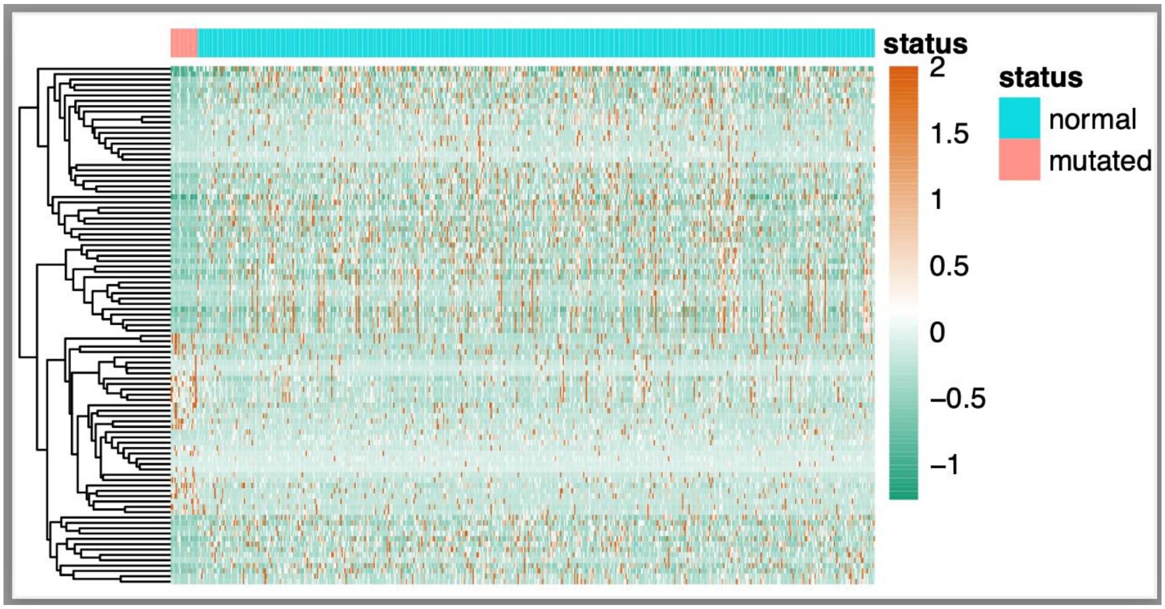


Figure 2: Heatmap of Intersected Differentially Expressed Genes

Enrichment Analysis

For the enrichment analysis we ran both the intersected DEGenes and the EdgeR only DEGenes. The intersected DEGenes enrichGO analysis had only 20 GO terms and all being in the BP category. The EdgeR DEGenes analysis had 228 GO term entries ranging across all 3 categories of GO terms.

This GO term chart represents, by color, the 3 categories: BP, CC, and MF. Figure 3 are the results from enrichGO using the intersected DEGenes. This graphic was created by SRPlot as well.

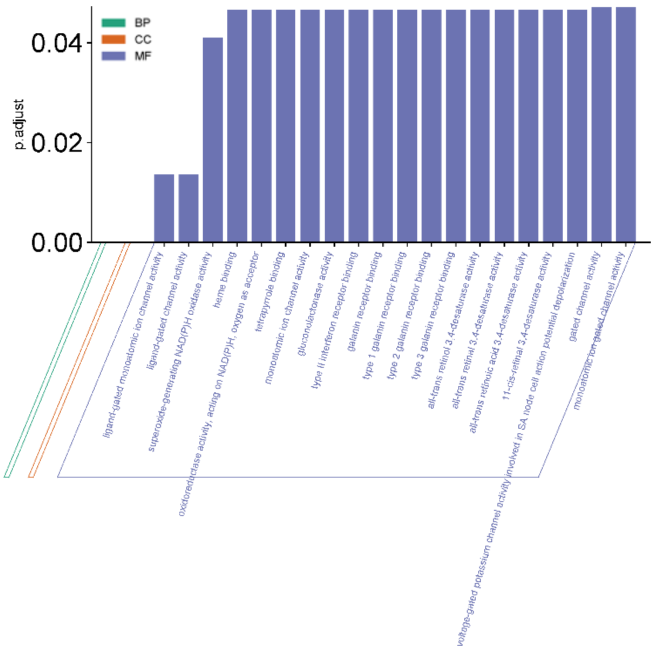


Figure 3: GO terms of Intersected DEGenes

This GO term chart represents, by color, the 3 categories: BP, CC, and MF. Figure 4 are the results from enrichGO using EdgeR DEGenes. This graphic was created using SRPlot.

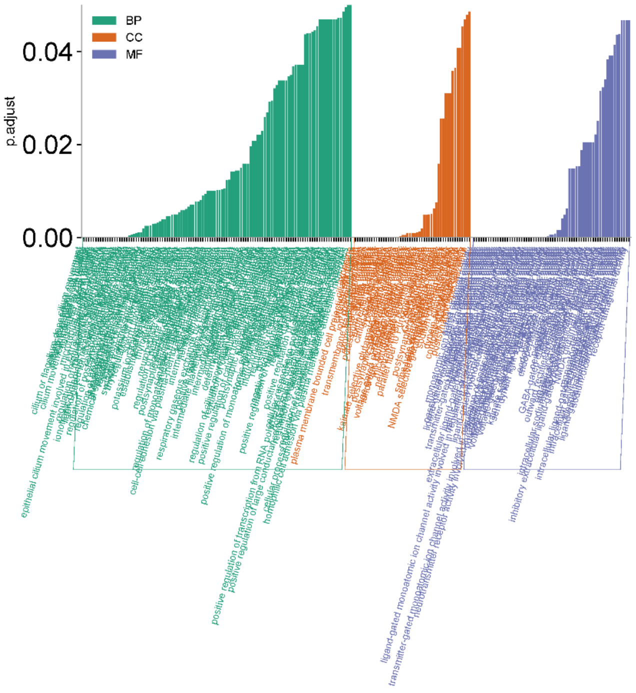


Figure 4: GO terms of EdgeR DEGenes

Associating Transcription Factors (TFs)

These are the top 10 transcription factors that exhibit the highest associations with the EdgeR DEGenes used as input in the web tool. The headers in Table 6 are explained in the methods section under 'Associating Transcription Factors (TFs)' in Table 4. The exported file from ChEA3 lists all 130 overlapping genes for each TF, providing detailed information instead of just numerical representation as seen in Table 6. Additionally, the web tool provides links to websites that explain the functions of the transcription factors and offers visualization tools to aid in result interpretation.

Table 6: Top 10 TF from ChEA3 Results from EdgeR DEGenes

Rank	TF	Score	Library	Overlapping Genes
1	CCDC17	2.5	ARCHS4 Coexpression,3;GTEx Coexpression,2	130
2	ZNF474	3.667	ARCHS4 Coexpression,1;Enrichr Queries,7;GTEx Coexpression,3	136
3	NEUROD4	25.67	ARCHS4 Coexpression,50;Enrichr Queries,1;GTEx Coexpression,26	87
4	SCRT1	30.5	ARCHS4 Coexpression,49;Enrichr Queries,12	56
5	PEG3	31.5	ARCHS4 Coexpression,27;GTEx Coexpression,36	58
6	LHX5	35.33	ARCHS4 Coexpression,45;Enrichr Queries,8;GTEx Coexpression,53	79
7	INSM1	47.33	ARCHS4 Coexpression,15;Enrichr Queries,104;GTEx Coexpression,23	72
8	POU3F3	47.67	ARCHS4 Coexpression,29;Enrichr Queries,83;GTEx Coexpression,31	72
9	SOX1	52.33	ARCHS4 Coexpression,125;Enrichr Queries,24;GTEx Coexpression,8	76
10	DACH2	53	ARCHS4 Coexpression,53;Enrichr Queries,71;GTEx Coexpression,35	64

Discussion

The choice to incorporate both Limma and EdgeR into our differential gene analysis workflow was driven by their respective advantages and disadvantages. EdgeR proves especially well-suited for RNAseq data, which aligns with our dataset. Moreover, it excels even when working with a small sample size (Robinson et al., 2010). In contrast, Limma offers a broad spectrum of analyses, including linear modeling and batch effect correction. Our dataset remained relatively uncomplicated, involving only a comparison between mutated and non-mutated groups, with a limited number of mutated patients (21 in total). Furthermore, Limma's original design caters more towards microarray data analysis rather than RNAseq data (Ritchie et al., 2015). The two tools also differ in their normalization methods and approaches to enhance statistical power (Ritchie et al., 2015). However, they share several similarities, including built-in quality control techniques, the use of standardized statistical methods, and their affiliation with the Bioconductor Project in R Studio. Running both analyses has provided us with a deeper comprehension of the DEGenes associated with FH, in line with the common practice among researchers who explore different analyses for various facets of their study. Notably, Limma yielded fewer DEGenes compared to EdgeR, possibly due to EdgeR's ExactTest being better suited to handle small sample sizes. Since EdgeR produced a more extensive list of DEGenes, we also subjected these genes to the enrichment analysis. The criteria for determining the significance of DEGenes remained consistent between both analyses. We employed a Log2FoldChange threshold of 1.58, a threshold used throughout our laboratory and by other researchers following similar workflows (Duan et al., 2022). That also facilitated the inclusion of more genes in the final list, broadening our exploration of potential connections to FH. Of the

DEGenes from the EdgeR list, the down-regulated genes has a lot more results compared to the up-regulated genes. This could be because when FH is mutated, there is a decrease in mitochondrial respiration. This means that there are many genes that are not functioning properly, some of which are not functioning at all (Schmidt et al., 2020a). After running both programs, we conducted an intersection analysis on the DEGenes to identify common genes between Limma and EdgeR. This cross-validation step aimed to ensure the attainment of the most reliable results. Out of all the DEGenes, only 97 exhibited overlap between Limma and EdgeR. The intersected genes, as well as the DEGenes list from EdgeR, were subsequently used for the enrichment analysis. The usage of two distinct files for the enrichment analysis allowed us to compare the similarities and disparities in the results.

For the enrichment analysis, we opted to use EnrichGO from the 'clusterProfiler' package in R Studio (T Wu et al., 2021). In this analysis, we conducted a comparative assessment between the intersected DEGenes and the results solely from the EdgeR DEG workflow. This approach was chosen because the EdgeR workflow yielded more DEGenes (compared to Limma), and the final results file generated from the EdgeR-only workflow closely aligned with our gene of interest.

As part of the EnrichGO workflow, the mapping of gene names to 'EntrezID' is conducted. However, there isn't always a 100% match between the input list of DEGenes and this was true for both of our datasets. Initially, we ran the intersected DEGenes, and the program indicated that 21% of the DEGenes failed to map to the 'EntrezIDs.' We set the p-value cutoff at 0.05, and the q-value cutoff at 0.1. The default q-value in EnrichGO is 0.2, but we tested both 0.1 and 0.2, and in the end, applying a filter of p.adjust < 0.05 produced the same results. Only 20 GO terms were identified from the intersected DEGenes, all within the 'molecular functions' ontology. Many of these GO terms were related to gated channels, galanin receptor binding, and oxidase activity.

Next, we analyzed the EdgeR DEGenes, where 18% failed to map to the EntrezIDs. The parameters and cutoffs for the EdgeR analysis matched those of the intersected genes analysis. We identified 228 GO terms, which were distributed as follows: 113 biological processes (BP), 49 cellular components (CC), and 66 molecular functions (MF). We created a barplot representing the top 10 categories based on 'Count,' depicted in Figure 5."

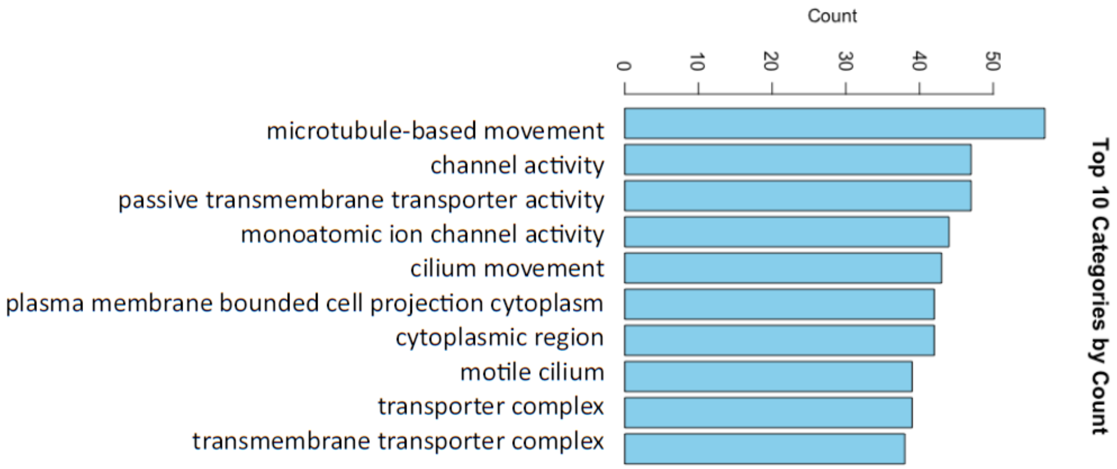


Figure 5: Top 10 Categories in EdgeR

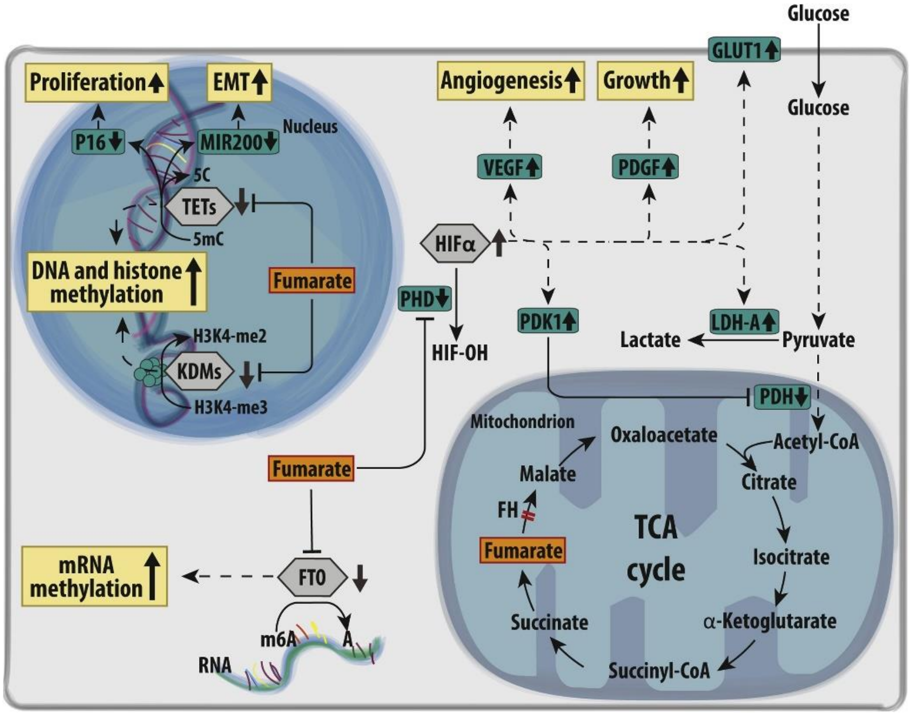


Figure 6: Oncogenic signalling mediated via aKGDDs inhibition in FH-deficient cells (Schmidt et al., 2020)

Based on Figure 5 and the results of the enrichment analysis, these findings suggest that FH is associated with numerous GO terms related to channel activity and transporter complexes. For instance, as shown in Figure 6 (Schmidt et al., 2020), pyruvate enters the mitochondria via the 'voltage-dependent anion channel,' which, within the Gene Ontology, falls under the 'channel activity' parent term (McCommis & Finck, 2015). We proceeded to compare the GO terms between the two runs.

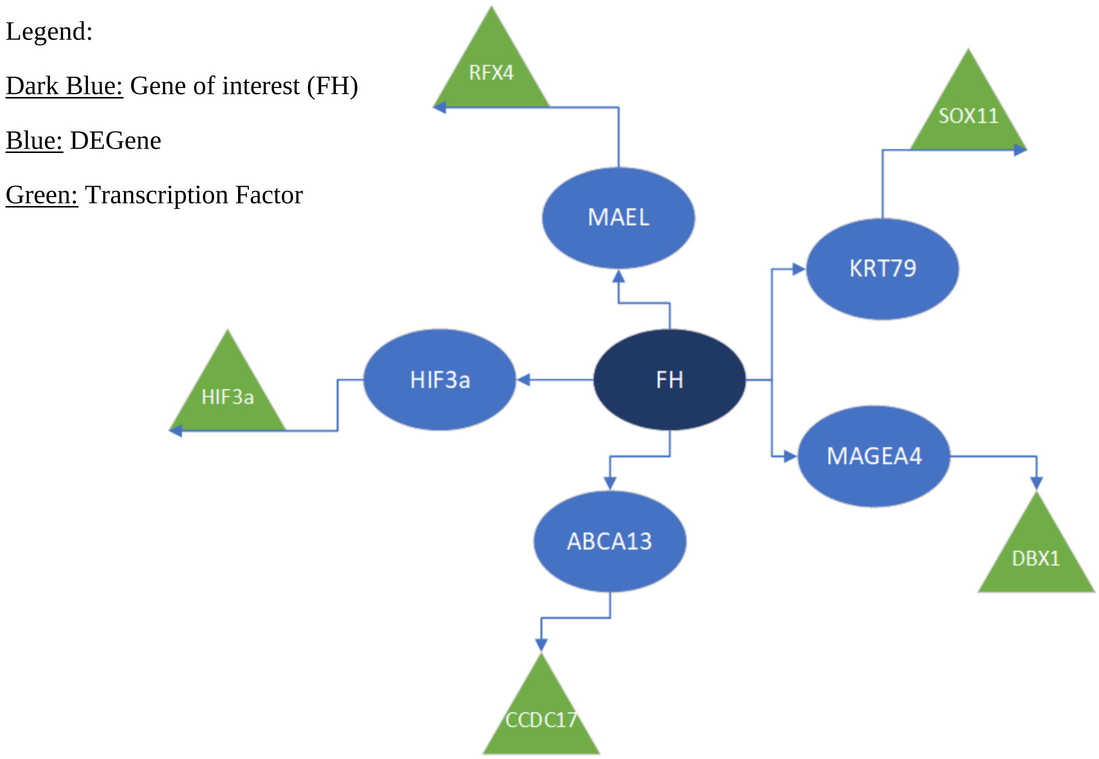
We found that there were 5 overlapping molecular function (MF) GO terms in the intersected DEGenes list. These 5 overlapping MF GO terms were related to gated channels, galanin receptor binding, and oxidase activity, which is also associated with the 'channel activity' term from the EdgeR-only run. Both of these connections further support the idea that our DEGene analysis and enrichment analysis were rigorous and generated pertinent results related to FH. It also indicates that the intersection of genes from the DEGene Analysis step, as opposed to using only the EdgeR results, was beneficial. These two distinct methodologies yielded similar results in the end.

Another tool that we tried but did not use the results from was Ingenuity Pathway Analysis from Qiagen (IPA). We were able to use the trial version of IPA but that limited our ability to the functions of IPA and we were only given two weeks to use the tool. The time frame was too small to learn everything about the tool and we were only allowed to export a certain number of files. These hinderances eventually prevented us from running and learning the full extent of the program.

Next, we proceeded to associate the transcription factors (TFs) related to the DEGenes. The purpose of associating the TFs with the DEGenes was to help create a diagram illustrating the relationship between FH, the DEGene, and the transcription factor. The ChEA3 webtool successfully linked all the DEGenes from the EdgeR analysis to their respective TFs. Likewise, we ran the intersected DEGenes through ChEA3 and obtained their results.

However, when we input the entire list of DEGenes into ChEA3, it alone did not provide significant insights for our research. Instead, upon reviewing other research papers related to FH, we became intrigued by the gene family associated with hypoxia-inducible factors (HIF) (Scagliola et al., 2019). According to Scagliola et al., 2019, 'FH loss-driven tumorigenesis has been associated with HIF-1a stabilization.' This piqued our interest to determine if the HIF gene was among the DEGenes in our analysis. In the EdgeR DEGene analysis, HIF3a was identified as a down-regulated DEGene. Although HIF3a did not appear in the EnrichGO analysis, it might be linked to the 'cytoplasmic region' depicted in Figure 5 since HIF primarily resides and functions in the cytoplasm.

The absence of HIF3a from the EnrichGO analysis could be attributed to its limited research compared to its other isoforms, HIF1a and HIF2a (Ravenna et al., 2016). HIF1a and HIF2a have been extensively documented in terms of their functions, downstream effects, and their associations with various diseases. While HIF3a and HIF1a did come up in the results from ChEA3, the HIF3a gene itself was not in the ‘overlapping genes’ column. This means that although the HIF3a gene was a DEGene in EdgeR, it was not considered significant in the ChEA3 webtool to be linked to the HIF3a TF. Figure 7 reflects our findings from DEGenes analysis and ChEA3 TF analysis. ABCA13 was one of the top genes associated with the top TF from ChEA3. MAGEA4 had the largest absolute value of log2FC in DEGene EdgeR analysis. KRT79 had the largest absolute value log2FC in the combined DEGene analysis. MAEL was the only gene that was present in the DEGene EdgeR analysis and the BioGRID website (Oughtred et al., 2020).



Legend:

Dark Blue: Gene of interest (FH)

Blue: DEGene

Green: Transcription Factor

Figure 7: FH Interaction Map

Conclusion

The purpose of this research paper was to gain a better understanding of Fumarate Hydratase and its impact on the Kreb Cycle when it is deficient. Following our differential gene analysis, enrichment analysis, and the association of transcription factors (TFs), our research suggests that the loss of FH has a downstream effect on the HIF3a TF and its corresponding gene, HIF3a. Notably, in line with other scholarly articles, we found that relatively little is known about the HIF3a gene. This knowledge gap presents an intriguing avenue for future research opportunities.

References

Goldman, M.J., Craft, B., Hastie, M. et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* 38, 675–678. https://doi.org/10.1038/s41587-020-0546-8

Matthew E. R., Belinda P., Di W., Yifang H., Charity W. L., Wei S., Gordon K. S. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47.https://doi.org/10.1093/nar/gkv007

Isaacs, S. J., Jung, J. Y., Mole, R. D., Lee, S., Torres-Cabala, C., Chung, Y., Merino, M., Trepel, J., Zbar, B., Toro, J., Ratcliffe, J. P., Linehan, M. W., Neckers, L. (2005). HIF overexpression correlates with biallelic loss of fumarate hydratase in renal cancer: Novel role of fumarate in regulation of HIF stability. *Cancer Cell*, 8(2), 143-153. https://doi.org/10.1016/j.ccr.2005.06.017.

Toro, J. R., Nickerson, M. L., Wei, M. H., Warren, M. B., Glenn, G. M., Turner, M. L., Stewart, L., Duray, P., Tourre, O., Sharma, N., Choyke, P., Stratton, P., Merino, M., Walther, M. M., Linehan, W. M., Schmidt, L. S., & Zbar, B. (2003). Mutations in the fumarate hydratase gene cause hereditary leiomyomatosis and renal cell cancer in families in North America. *American journal of human genetics*, 73(1), 95–106. https://doi.org/10.1086/376435

Schmidt, C., Sciacovelli, M., Frezza, C. (2020). Fumarate hydratase in cancer: A multifaceted tumour suppressor. *Seminars in Cell & Developmental Biology*, 98, 15-25. https://doi.org/10.1016/j.semcdb.2019.05.002.

Fuchs, T. L., Luxford, C., Clarkson, A., Sheen, A., Sioson, L., Elston, M., Croxson, M. S., Dwight, T., Benn, D. E., Tacon, L., Field, M., Ahadi, M. S., Chou, A., Clifton-Bligh, R. J., & Gill, A. J. (2023). A Clinicopathologic and Molecular Analysis of Fumarate Hydratase-deficient

Pheochromocytoma and Paraganglioma. *The American journal of surgical pathology*, 47(1), 25–36. https://doi.org/10.1097/PAS.0000000000001945

Wang, G., Wang, F., Huang, Q., Li, Y., Liu, Y., & Wang, Y. (2015). Understanding Transcription Factor Regulation by Integrating Gene Expression and DNase I Hypersensitive Sites. *BioMed research international*, 2015, 757530. https://doi.org/10.1155/2015/757530

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (Oxford, England), 26(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*, 16(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1), 25–29. https://doi.org/10.1038/75556

Keenan, A. B., Torre, D., Lachmann, A., Leong, A. K., Wojciechowicz, M. L., Utti, V., Jagodnik, K. M., Kropiwnicki, E., Wang, Z., & Ma'ayan, A. (2019). ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic acids research*, 47(W1), W212–W224. https://doi.org/10.1093/nar/gkz446

Balakrishnan, R., Harris, M. A., Huntley, R., Van Auken, K., & Cherry, J. M. (2013). A guide to best practices for Gene Ontology (GO) manual annotation. *Database: the journal of biological databases and curation*, 2013, bat054. https://doi.org/10.1093/database/bat054

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., & Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*, 6(269), pl1. https://doi.org/10.1126/scisignal.2004088

de Bruijn, I., Kundra, R., Mastrogiacomo, B., Tran, T. N., Sikina, L., Mazor, T., Li, X., Ochoa, A., Zhao, G., Lai, B., Abeshouse, A., Baiceanu, D., Ciftci, E., Dogrusoz, U., Dufilie, A., Erkoc, Z., Garcia Lara, E., Fu, Z., Gross, B. E., Haynes, C. D., … Schultz, N. (2023). Analysis and Visualization of Longitudinal Genomic and Clinical Data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. *Cancer research*, 10.1158/0008-5472.CAN-23-0816. Advance online publication. https://doi.org/10.1158/0008-5472.CAN-23-0816

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., Zhu, J., & Haussler, D. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nature biotechnology*, 38(6), 675–678. https://doi.org/10.1038/s41587-020-0546-8

Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., & Smyth, G. K. (2016). ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE

GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION. *The annals of applied statistics*, 10(2), 946–963. https://doi.org/10.1214/16-AOAS920

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Chen, Y., Lun, A. T., & Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, 5, 1438. https://doi.org/10.12688/f1000research.8987.2

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* (Cambridge (Mass.)), 2(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

McCommis, K. S., & Finck, B. N. (2015). Mitochondrial pyruvate transport: a historical perspective and future research directions. *The Biochemical journal*, 466(3), 443–454. https://doi.org/10.1042/BJ20141171

Scagliola, A., Mainini, F., & Cardaci, S. (2020). The Tricarboxylic Acid Cycle at the Crossroad Between Cancer and Immunity. *Antioxidants & redox signaling*, 32(12), 834–852. https://doi.org/10.1089/ars.2019.7974

Ravenna, L., Salvatori, L., & Russo, M. A. (2016). HIF3α: the little we know. *The FEBS journal*, 283(6), 993–1003. https://doi.org/10.1111/febs.13572

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B. J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., & Tyers, M. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein science : a publication of the Protein Society*, 30(1), 187–200. https://doi.org/10.1002/pro.3978

Duan, S., Xu, Z., Li, X. Y., Liao, P., Qin, H. K., Mao, Y. P., Dai, W. S., Ma, H. J., & Bao, M. L. (2022). Dodder-transmitted mobile systemic signals activate a salt-stress response characterized by a transcriptome change in Citrus sinensis. Frontiers in plant science, 13, 986365. https://doi.org/10.3389/fpls.2022.986365

Vita

Sydney Lac attended Manvel High School in Manvel, Texas, graduating in May 2017. Subsequently, she pursued her Bachelor of Science and Arts degree at the University of Texas at Austin, majoring in Biochemistry with a minor in Business. Following her graduation from UT Austin in May 2020, she joined Baylor Genetics and maintained her position while concurrently pursuing graduate studies. In August 2021, she commenced her graduate studies at The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.